BERT Knows Punta Cana is not just Beautiful, it's Gorgeous : Ranking Scalar Adjectives with Contextualised Representations

Gari Soler, Aina

The Association for Computational Linguistics
2020

Gari Soler , A & Apidianaki , M 2020 , BERT Knows Punta Cana is not just Beautiful, it's Gorgeous : Ranking Scalar Adjectives with Contextualised Representations . in B Webber , T Cohn , Y He & Y Liu (eds) , Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) . The Association for Computational Linguistics , þÿ S t r o u d s b u r g , p p . 7 3 7 1 7 3 8 5 , T h e 2 0 2 0 C o n f e r e n c e o n E m p i r i c a l M e t h Language Processing , 16/11/2020 . https://doi.org/10.18653/v1/2020.emnlp-main.598

http://hdl.handle.net/10138/326108
https://doi.org/10.18653/v1/2020.emnlp-main.598

BERT Knows Punta Cana is not just *beautiful*, it's *gorgeous*: Ranking Scalar Adjectives with Contextualised Representations

Aina Garí Soler
Université Paris-Saclay
CNRS, LIMSI
91400 Orsay, France
aina.gari@limsi.fr

Marianna Apidianaki
Department of Digital Humanities
University of Helsinki
Helsinki, Finland
marianna.apidianaki@helsinki.fi

Abstract

Adjectives like *pretty*, *beautiful* and *gorgeous* describe positive properties of the nouns they modify but with different intensity. These differences are important for natural language understanding and reasoning. We propose a novel BERT-based approach to intensity detection for scalar adjectives. We model intensity by vectors directly derived from contextualised representations and show they can successfully rank scalar adjectives. We evaluate our models both intrinsically, on gold standard datasets, and on an Indirect Question Answering task. Our results demonstrate that BERT encodes rich knowledge about the semantics of scalar adjectives, and is able to provide better quality intensity rankings than static embeddings and previous models with access to dedicated resources.

1 Introduction

Scalar adjectives describe a property of a noun at different degrees of intensity. Identifying the scalar relationship that exists between their meaning (for example, the increasing intensity between *pretty*, *beautiful* and *gorgeous*) is useful for text understanding, for both humans and automatic systems. It can serve to define the sentiment and subjectivity of a text, perform inference and textual entailment (Van Tiel et al., 2016; McNally, 2016), build question answering and recommendation systems (de Marneffe et al., 2010), and assist language learners in distinguishing between semantically similar words (Sheinman and Tokunaga, 2009).

We investigate the knowledge that the pretrained BERT model (Devlin et al., 2019) encodes about the intensity expressed on an adjective scale. Given that this property is acquired by humans during language learning, we expect a language model (LM) exposed to massive amounts of text data during training to have also acquired some notion of adjective intensity. In what follows, we explore this hypothesis using representations extracted from different layers of this deep neural model. Since the scalar relationship between adjectives is context-dependent (Kennedy and McNally, 2005) (e.g., what counts as *tall* may vary from context to context), we consider the contextualised representations produced by BERT to be a good fit for this task. We also propose a method inspired by gender bias work (Bolukbasi et al., 2016; Dev and Phillips, 2019) for detecting the intensity relationship of two adjectives on the fly. We view intensity as a direction in the semantic space which, once identified, can serve to determine the intensity of new adjectives.

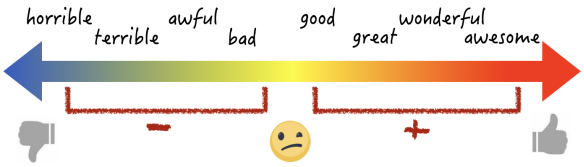

Figure 1: Full scale of adjectives describing positive and negative sentiment at different degrees from the SO-CAL dataset (Taboada et al., 2011).

Our work falls in the neural network interpretation paradigm which explores the knowledge about language encoded in the representations of deep learning models (Voita et al., 2019a; Clark et al., 2019; Voita et al., 2019b; Tenney et al., 2019; Talmor et al., 2019). The bulk of this interpretation work addresses structural aspects of language such as syntax, word order, or number agreement (Linzen et al., 2016; Hewitt and Manning, 2019; Hewitt and Liang, 2019; Rogers et al., 2020); shallow semantic phenomena closely related to syntax such as semantic role labelling and coreference (Tenney et al., 2019; Kovaleva et al., 2019); or the symbolic reasoning potential of language model representations (Talmor et al., 2019). Our

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7371–7385,
November 16–20, 2020.

work makes a contribution towards the study of the knowledge pre-trained LMs encode about word meaning, generally overlooked until now in interpretation work.

We evaluate the representations generated by BERT against gold standard adjective intensity estimates (de Melo and Bansal, 2013; Wilkinson, 2017; Cocos et al., 2018) and apply them directly to a question answering task (de Marneffe et al., 2010). Our results show that BERT clearly encodes the intensity variation between adjectives on scales describing different properties. Our proposed method can be easily applied to new datasets and languages where scalar adjective resources are not available.[1]

2 Related Work

The analysis of scalar adjective relationships in the literature has often been decomposed into two steps: Grouping related adjectives together and ranking adjectives in the same group according to intensity. The first step can be performed by distributional clustering approaches (Hatzivassiloglou and McKeown, 1993; Pang et al., 2008) which can also address adjectival polysemy. *Hot*, for example, can be on the TEMPERATURE scale (a *warm* → *hot* → *scalding* drink), the ATTRACTIVENESS (a *pretty* → *hot* → *sexy* person) or the INTEREST scale (an *interesting* → *hot* topic), depending on the attribute it modifies.

Other works (Sheinman and Tokunaga, 2009; de Melo and Bansal, 2013; Wilkinson, 2017) directly address the second step, ranking groups of semantically related adjectives from lexicographic resources (e.g., WordNet) (Fellbaum, 1998). This ranking is the focus of this work. We show that BERT contextualised representations encode rich information about adjective intensity, and can provide high quality rankings of adjectives in a scale.

Adjective ranking has been traditionally performed using pattern-based approaches which extract lexical or syntactic patterns indicative of an intensity relationship from large corpora (Sheinman and Tokunaga, 2009; de Melo and Bansal, 2013; Sheinman et al., 2013; Shivade et al., 2015). For example, the patterns "*X, but not Y*" and "*not just X but Y*" provide evidence that X is an adjective less intense than Y. Another common approach is lexicon-based and draws upon a resource that maps adjectives to scores encoding sentiment polarity (positive or negative) and intensity. Such resources can be manually created, like the SO-CAL lexicon (Taboada et al., 2011), or automatically compiled by mining adjective orderings from star-valued product reviews where people's comments have associated ratings (de Marneffe et al., 2010; Rill et al., 2012; Sharma et al., 2015; Ruppenhofer et al., 2014). Cocos et al. (2018) combine knowledge from lexico-syntactic patterns and the SO-CAL lexicon with paraphrases in the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013; Pavlick et al., 2015).

Our approach is novel in that it does not need specified patterns or access to lexicographic resources. It, instead, relies on the knowledge about intensity encoded in scalar adjectives' contextualised representations. Our best performing method is inspired by work on gender bias which relies on simple vector arithmetic to uncover gender-related stereotypes. A gender direction is determined (for example, by comparing the embeddings of *she* and *he*, or *woman* and *man*) and the projection of the vector of a potentially biased word on this direction is then calculated (Bolukbasi et al., 2016; Zhao et al., 2018). We extend this method to scalar adjectives and BERT representations.

Kim and de Marneffe (2013) also consider vector distance in the semantic space to encode scalar relationships between adjectives. They specifically examine a small set of word pairs, and observe that the middle point in space between the word2vec (Mikolov et al., 2013) embeddings of two antonyms (e.g., *furious* and *happy*) falls close to the embedding of a mid-ranked word in their scale (e.g., *unhappy*). Their experiments rely on antonym pairs extracted from WordNet. We show that contextualised representations are a better fit for this task than static embeddings, encoding rich information about adjectives' meaning and intensity.

3 Data

We experiment with three scalar adjective datasets.

DEMELO (de Melo and Bansal, 2013).[2] Adjective sets were extracted from WordNet 'dumbbell' structures (Gross and Miller, 1990). The sets represent full-scales (e.g., from *horrible* to *awesome*) and are partitioned into half-scales (from *horrible* to *bad*, and from *good* to *awesome*) based on pattern-based evidence in the Google N-Grams cor-

[1] Our code and data are available at https://github.com/ainagari/scalar_adjs

[2] http://demelo.org/gdm/intensity/

Dataset	Adjective scale
DEMELO	[*soft* → *quiet* → *inaudible* → *silent*] [*thick* → *dense* → *impenetrable*]
CROWD	[*fine* → *remarkable* → *spectacular*] [*scary* || *frightening* → *terrifying*]
WILKINSON	[*damp* → *moist* → *wet*] [*dumb* → *stupid* → *idiotic*]

Table 1: Examples of scales in each dataset. '||' denotes a tie between adjectives of the same intensity.

pus (Brants and Franz, 2006). The dataset contains 87 half-scales with 548 adjective pairs, manually annotated for intensity relations ($<$, $>$, and $=$).

CROWD (Cocos et al., 2018).[3] The dataset consists of a set of adjective scales with high coverage of the PPDB vocabulary. It was constructed by a three-step process: Crowd workers were first asked to determine whether pairs of adjectives describe the same attribute (e.g., TEMPERATURE) and should, therefore, belong to the same scale. Sets of same-scale adjectives were then refined over multiple rounds. Finally, workers ranked the adjectives in each set by intensity. The final dataset includes 330 adjective pairs along 79 half-scales.

WILKINSON (Wilkinson and Oates, 2016).[4] This dataset was generated through crowdsourcing. Crowd workers were presented with small seed sets (e.g., *huge, small, microscopic*) and were asked to propose similar adjectives, resulting in twelve adjective sets. Sets were automatically cleaned for consistency, and then annotated for intensity by the crowd workers. The original dataset contains full scales. We use its division in 21 half-scales (with 61 adjective pairs) proposed by Cocos et al. (2018).

In the rest of the paper, we use the term "scale" to refer to the half-scales contained in these datasets. Table 1 shows examples from each one of them.

4 BERT Contextualised Representations

4.1 Sentence Collection

To explore the knowledge BERT has about relationships in an adjective scale s, we generate a contextualised representation for each $a \in s$ in the same context. Since such cases are rare in running text, we construct two sentence sets that satisfy this condition using the ukWaC corpus (Baroni et al., 2009)[5] and the Flickr 30K dataset (Young et al., 2014).[6] For every $s \in D$, a dataset from Section 3, and for each $a \in s$, we collect 1,000 instances (sentences) from each corpus.[7] We substitute each instance i of $a \in s$, with each $b \in s$ where $b \neq a$, creating $|s| - 1$ new sentences.[8] For example, for an instance of *thick* from the scale [*thick* → *dense* → *impenetrable*] in Table 1, we generate two new sentences where *thick* is substituted by each of the other adjectives in the same context.

4.2 Sentence Cleaning

Hearst patterns We filter out sentences where substitution should not take place, such as cases of specialisation or instantiation. In this way, we avoid replacing *deceptive* with *fraudulent* and *false* in sentences like "*Viruses and other deceptive software*", "*Deceptive software such as viruses*", "*Deceptive software, especially viruses*".[9] We parse the sentences with `stanza` (Qi et al., 2020) to reveal their dependency structure, and use Hearst lexico-syntactic patterns (Hearst, 1992) to identify sentences describing *is-a* relationships between nouns in a text. More details about this filtering are given in Appendix A.

Language Modelling criteria Adjectives that belong to the same scale might not be replaceable in all contexts. Polysemy can also influence their substitutability (e.g., *warm weather* is a bit *hot*, but a *warm smile* is *friendly*). In order to select contexts where $\forall a \in s$ fit, we measure the fluency of the sentences generated through substitution. We use a score assigned to each sentence by context2vec (Melamud et al., 2016) which reflects how well an $a \in s$ fits a context by measuring the cosine similarity between a and the context representation. We also experimented with calculating the

[3] https://github.com/acocos/scalar-adj

[4] https://github.com/Coral-Lab/scales

[5] http://u.cs.biu.ac.il/~nlp/resources/downloads/context2vec/

[6] Flickr contains crowdsourced captions for 31,783 images describing everyday activities, events and scenes. We consider objective descriptions to be a better fit for our task than subjective statements, which might contain emphatic markers. For example, *impossible* would be a bad substitute for *impractical* in the sentence "*What you ask for is too impractical*".

[7] ukWaC has perfect coverage. Flickr 30K covers 96.56% of the DEMELO scales and 86.08% of the CROWD scales. A scale s is not covered when no $a \in s$ is found in a corpus.

[8] We make a minor adjustment of the substituted data by replacing the indefinite article a with an when the adjective that follows starts with a vowel, and the inverse when it starts with a consonant.

[9] This would especially be a problem when considering adjectives with different polarity on a full scale (e.g., *deceptive* and *honest*).

perplexity assigned by BERT to a sentence generated through substitution, and with replacing the original a instance with the [MASK] token and getting the BERT probability for each $a \in s$ as a filler for that slot. context2vec was found to make better substitutability estimates.[10]

We use a 600-dimensional context2vec model in our experiments, pre-trained on ukWaC.[11] We calculate the context2vec score for all sentences generated for a scale s through substitution, and keep the ten with the lowest standard deviation (STD). Low STD for a sentence means that $\forall a \in s$ are reasonable choices in this context. For comparison, we also randomly sample ten sentences from all the ukWaC sentences collected for each scale. We call the sets of sentences ukWaC, Flickr and Random SENT-SETs.

We extract the contextualised representation for each $a \in s$ in the ten sentences retained for scale s, using the pre-trained `bert-base-uncased` model.[12] This results in $|s| * 10$ BERT representations for each scale. We repeat the procedure for every BERT layer. Examples of the obtained sentences are given in Appendix B.

5 Scalar Adjectives Ranking

5.1 Ranking with a Reference Point

In our first ranking experiment, we explore whether BERT encodes adjective intensity relative to a reference point, that is the adjective with the highest intensity (a_{ext}) in a scale s.

Method We rank $\forall a \in s$ where $a \neq a_{ext}$ by intensity by measuring the cosine similarity between their representation and that of a_{ext} in the ten ukWaC sentences retained for s, and in every BERT layer. For example, to rank [*pretty, beautiful, gorgeous*] we measure the similarity of the representations of *pretty* and *beautiful* to that of *gorgeous*. We then average the similarities obtained for each a and use these values for ranking. We refer to this method as BERTSIM.

We evaluate the quality of the ranking for a scale by measuring its correlation with the gold standard ranking in the corresponding dataset D using Kendall's τ and Spearman's ρ correlation coefficients.[13] We also measure the model's pairwise accuracy (P-ACC) which shows whether it correctly predicted the relative intensity ($<$, $>$, $=$) for each pair a_i-$a_j \in s$ with $i \neq j$. During evaluation, we do not take into account scales where only one adjective is left ($|s| = 1$) after removing a_{ext} (26 out of 79 scales in CROWD; 9 out of 21 scales in WILKINSON).

Baselines We compare the BERTSIM method to two baselines which rank adjectives by frequency (FREQ) and number of senses (SENSE). We make

Dataset	Metric	BERTSIM	FREQ	SENSE
DEMELO	P-ACC	**0.591**$_{11}$	0.571	0.493
	τ	**0.364**$_{11}$	0.304	0.192
	ρ_{avg}	**0.389**$_{11}$	0.309	0.211
CROWD	P-ACC	**0.646**$_{11}$	0.608	0.570
	τ	**0.498**$_{11}$	0.404	0.428
	ρ_{avg}	0.494$_{11}$	0.499	**0.537**
WILKINSON	P-ACC	**0.913**$_{9}$	0.739$_{9}$	0.739$_{9}$
	τ	**0.826**$_{9}$	0.478	0.586
	ρ_{avg}	**0.724**$_{9}$	0.345	0.493

Table 2: BERTSIM results on each dataset using contextualised representations from the ukWaC SENT-SET. Subscripts denote the best-performing BERT layer.

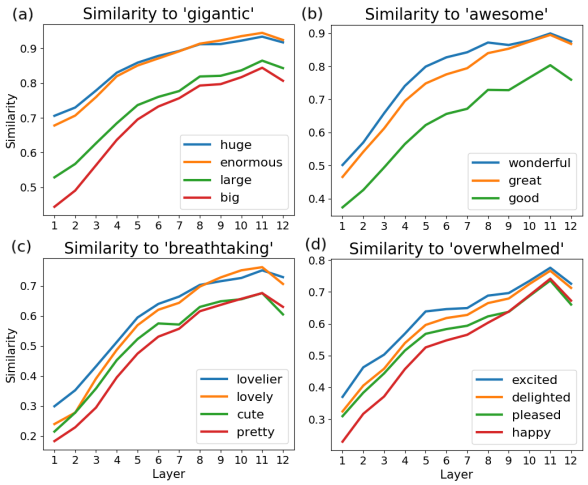

Figure 2: Examples of BERTSIM ranking predictions across layers using ukWaC sentences for four adjective scales: (a) [*big* → *large* → *enormous* → *huge* → *gigantic*], (b) [*good* → *great* → *wonderful* → *awesome*], (c) [*cute* → *pretty* → *lovely* → *lovelier* → *breathtaking*], (d) [*pleased* → *happy* → *excited* → *delighted* → *overwhelmed*]. (a) and (b) are from WILKINSON, (c) and (d) are from CROWD.

[10] We use as development set for this exploration a sample of 500 sentence pairs from the Concepts in Context (CoInCo) corpus (Kremer et al., 2014) that we will share along with our code. Details on the constitution of this sample are in Appendix B.

[11] http://u.cs.biu.ac.il/~nlp/resources/downloads/context2vec/

[12] When an adjective is split into multiple wordpieces (Wu et al., 2016), we average them to obtain its representation.

[13] We report correlations as a weighted average using the number of adjective pairs in a scale as weights.

the assumption that words with low intensity (e.g., *good, old*) are more frequent and polysemous than their extreme counterparts on the same scale (e.g., *awesome, ancient*). This assumption relies on the following two intuitions which we empirically validate: (a) Extreme adjectives tend to restrict the denotation of a noun to a smaller class of referents than low intensity adjectives (Geurts, 2010). We hypothesise that extreme adjectives denote more exceptional and less frequently encountered properties of nouns than low intensity adjectives on the same scale. This is also reflected in the directionality of their entailment relationship (e.g., *awesome* $\rightarrow$ *good*, *good* $\not\rightarrow$ *awesome*); low intensity adjectives should thus be more frequently encountered in texts. We test this assumption using frequency counts in Google Ngrams (Brants and Franz, 2006), and find that the least intense adjective is indeed more frequent than the most extreme adjective in 75% of the scales; (b) Since frequent words tend to be more polysemous (Zipf, 1945), we also expect that low intensity adjectives would have more senses than extreme ones. This is confirmed by their number of senses in WordNet: in 67% of the scales, the least intense adjective has a higher number of senses than its extreme counterpart.

Results We present the results of this evaluation in Table 2. Overall, similarities derived from BERT representations encode well the notion of intensity, as shown by the moderate to high accuracy and correlation in the three datasets. The good results obtained by the FREQ and SENSE baselines (especially on CROWD) highlight the relevance of frequency and polysemy for scalar adjective ranking, and further validate our assumptions.

Figure 2 shows ranking predictions made by BERTSIM in different layers of the model. Predictions are generally stable and reasonable across layers, despite not always being correct. For example, the similarly-intense *happy* and *pleased* are inverted in some layers but are not confused with adjectives further up the scale (*excited, delighted*). Note that *happy* and *pleased* are in adjacent positions in the CROWD ranking, and form a tie in the DEMELO dataset.

5.2 Ranking without Specified Boundaries

In real life scenarios, scalar adjective interpretation is performed without concrete reference points (e.g., a_{ext}). We need to recognize that a *great book* is better than a *well-written* one, without necessarily detecting their relationship to *brilliant*.

Method Our second adjective ranking method draws inspiration from word analogies in gender bias work, where a gender subspace is identified in word-embedding space by calculating the main direction spanned by the differences between vectors of gendered word pairs (e.g., $\overrightarrow{he}$ - $\overrightarrow{she}$, $\overrightarrow{man}$ - $\overrightarrow{woman}$) (Bolukbasi et al., 2016; Dev and Phillips, 2019; Ravfogel et al., 2020; Lauscher et al., 2020).

We propose to obtain an *intensity direction* by subtracting the representation of a mild intensity adjective a_{mild} from that of an extreme adjective a_{ext} on the same scale. By subtracting *pretty* from *gorgeous*, for example, which express a similar core meaning (they are both on the BEAUTY scale) but with different intensity, we expect the resulting $\overrightarrow{dVec} = \overrightarrow{gorgeous}$ - $\overrightarrow{pretty}$ embedding to represent this notion of intensity (or degree). We can then compare other adjectives' representations to $\overrightarrow{dVec}$, and rank them according to their cosine similarity[14] to this intensity vector: the closer an adjective is to $\overrightarrow{dVec}$, the more intense it is.

We calculate the $\overrightarrow{dVec}$ for each $s \in D$ (a dataset from Section 3) using the most extreme (a_{ext}) and the mildest (a_{mild}) words in s. We experiment with BERT embeddings from the SENT-SETs generated through substitution as described in Section 4, and with static word2vec embeddings (Mikolov et al., 2013) trained on Google News.[15] We build a $\overrightarrow{dVec}$ from every sentence (context) c in the set of ten sentences C for a scale s by subtracting the BERT representation of a_{mild} in c from that of a_{ext} in c. We average the ten $\overrightarrow{dVec}$'s obtained for s and construct a global $\overrightarrow{dVec}$ for the dataset D by averaging the vectors of $\forall s \in D$. For a fair evaluation, we perform a lexical split in the data used for deriving $\overrightarrow{dVec}$ and the data used for testing. When evaluating on CROWD, we calculate a $\overrightarrow{dVec}$ vector on DEMELO (DIFFVEC-DM) and one on WILKINSON (DIFFVEC-WK), omitting all scales where a_{ext} or a_{mild} are present in CROWD. We do the same for the other datasets.

To obtain the $\overrightarrow{dVec}$ of a s with static embeddings, we simply calculate the difference between the word2vec embeddings of a_{ext} and a_{mild} in s.

Results For evaluation, we use the same metrics as in Section 5.1. We compare our results to the

[14] We also tried the dot product of the vectors. The results were highly similar to the ones obtained using the cosine.

[15] We use the `magnitude` library (Patel et al., 2018).

			DEMELO (DM)			CROWD (CD)			WILKINSON (WK)		
		Method	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}
BERT	ukWaC	DIFFVEC-DM	-	-	-	**0.739**$_{12}$	**0.674**$_{12}$	**0.753**$_{12}$	0.918$_{6}$	0.836$_{6}$	0.839$_{6}$
		DIFFVEC-CD	0.646$_{8}$	0.431$_{8}$	**0.509**$_{8}$	-	-	-	0.869$_{11}$	0.738$_{11}$	0.829$_{11}$
		DIFFVEC-WK	0.584$_{9}$	0.303$_{9}$	0.313$_{10}$	0.706$_{10}$	0.603$_{9}$	0.687$_{9}$	-	-	-
	Flickr	DIFFVEC-DM	-	-	-	0.730$_{12}$	0.667$_{12}$	0.705$_{10}$	**0.934**$_{9}$	**0.869**$_{9}$	**0.871**$_{9}$
		DIFFVEC-CD	0.620$_{10}$	0.377$_{10}$	0.466$_{10}$	-	-	-	0.902$_{7}$	0.803$_{7}$	0.798$_{7}$
		DIFFVEC-WK	0.579$_{1}$	0.294$_{1}$	0.321$_{1}$	0.702$_{8}$	0.608$_{8}$	0.677$_{8}$	-	-	-
	Random	DIFFVEC-DM	-	-	-	**0.739**$_{12}$	0.673$_{12}$	0.743$_{12}$	0.918$_{6}$	0.836$_{6}$	0.839$_{6}$
		DIFFVEC-CD	0.626$_{8}$	0.388$_{8}$	0.466$_{8}$	-	-	-	0.836$_{12}$	0.672$_{12}$	0.790$_{10}$
		DIFFVEC-WK	0.557$_{9}$	0.246$_{9}$	0.284$_{6}$	0.703$_{8}$	0.598$_{8}$	0.676$_{8}$	-	-	-
word2vec		DIFFVEC-DM	-	-	-	0.657	0.493	0.543	0.787	0.574	0.663
		DIFFVEC-CD	0.633	0.398	0.444	-	-	-	0.803	0.607	0.637
		DIFFVEC-WK	0.593	0.323	0.413	0.618	0.413	0.457	-	-	-
Baseline		FREQ	0.575	0.271	0.283	0.606	0.386	0.452	0.754	0.508	0.517
		SENSE	0.493	0.163	0.165	0.658	0.498	0.595	0.721	0.586	0.575
		Cocos et al. '18	**0.653**	**0.633**	-	0.639	0.495	-	0.754	0.638	-

Table 3: Results of our DIFFVEC adjective ranking method on the DEMELO, CROWD, and WILKINSON datasets. We report results with contextualised (BERT) representations obtained from different SENT-SETs (ukWaC, Flickr, Random) and with static (word2vec) vectors. We compare to the frequency (FREQ) and number of senses (SENSE) baselines, and to results from previous work (Cocos et al., 2018). Results for a dataset are missing (-) when the dataset was used for building the $\overrightarrow{dVec}$ intensity vector.

FREQ and SENSE baselines, and to the best results obtained by Cocos et al. (2018) who use information obtained from lexico-syntactic patterns, a lexicon annotated with intensity (SO-CAL) (Taboada et al., 2011), and paraphrases from PPDB.[16] Results are presented in Table 3. The DIFFVEC method gets remarkably high performance compared to previous results, especially when $\overrightarrow{dVec}$ is calculated with BERT embeddings. With the exception of Kendall's τ and pairwise accuracy on the DEMELO dataset, DIFFVEC outperforms results from previous work and the baselines across the board. We believe the lower correlation scores on the DEMELO dataset to be due to the large amount of ties present in this dataset: 44% of scales in DEMELO contain ties, versus 30% in CROWD and 0% in WILKINSON, where we obtain better results. Our models cannot easily predict ties using similarities which are continuous values. To check whether our assumption is correct, we make a simple adjustment to DIFFVEC so that it can propose ties if the vectors of two adjectives are similarly close to $\overrightarrow{dVec}$. Overall, this results in a small decrease in pairwise accuracy and a slight increase in correlation in DEMELO and CROWD. Complete results of this additional evaluation are given in Appendix C.

[16] We do not report Spearman's ρ from Cocos et al. (2018) because it was calculated differently: They measure it a single time for each dataset, treating each adjective as a single data point.

The composition of the SENT-SETs used for building BERT representations also plays a role on model performance. Overall, the selection method described in Section 4 offers a slight advantage over random selection, with ukWaC and Flickr sentences improving performance on different datasets. Note, however, that results for Flickr are calculated on the scales for which sentences were available (96.56% of DEMELO scales and 86.08% from CROWD).

The best-performing BERT layers are generally situated in the upper half of the Transformer network. The only exception is DIFFVEC-WK with the Flickr SENT-SET on DEMELO, where all layers perform similarly. The FREQ and SENSE baselines get lower performance than our method with BERT embeddings. SENSE manages to give results comparable to DIFFVEC with static embeddings and to previous work (Cocos et al., 2018) in one dataset (CROWD), but is still outperformed by DIFFVEC with contextualised representations.

We can also compare our results to those obtained by a purely pattern-based method on the same datasets, reported by Cocos et al. (2018). This method performs well on DEMELO ($\tau = 0.663$) because of its high coverage on this dataset, which was compiled by finding adjective pairs that also match lexical patterns. The performance of the pattern-based method is much lower than that of our models in the other two datasets ($\tau = 0.203$ on CROWD, $\tau = 0.441$ on WILKINSON), and its

		DEMELO			
		# Scales	P-ACC	τ	ρ_{avg}
BERT	ukWaC	1 (+)	0.653_{9}	0.438_{9}	0.489_{11}
		1 (−)	0.611_{10}	0.350_{10}	0.424_{11}
		5	0.650_{10}	0.430_{10}	0.514_{10}
	Flickr	1 (+)	0.656_{8}	0.449_{8}	0.504_{8}
		1 (−)	0.600_{3}	0.324_{3}	0.375_{5}
		5	0.647_{12}	0.426_{12}	0.498_{11}
	Random	1 (+)	$\mathbf{0.659}_{11}$	$\mathbf{0.451}_{11}$	0.493_{11}
		1 (−)	0.608_{12}	0.340_{12}	0.421_{10}
		5	0.653_{11}	0.442_{11}	$\mathbf{0.538}_{10}$
word2vec		1 (+)	0.602	0.334	0.364
		1 (−)	0.613	0.359	0.412
		5	0.641	0.415	0.438

		CROWD			
		# Scales	P-ACC	τ	ρ_{avg}
BERT	ukWaC	1 (+)	$\mathbf{0.709}_{12}$	$\mathbf{0.611}_{12}$	$\mathbf{0.670}_{12}$
		1 (−)	0.648_{10}	0.477	0.507_{10}
		5	0.700_{11}	0.595_{10}	0.673_{10}
	Flickr	1 (+)	0.676_{12}	0.552_{8}	0.612_{8}
		1 (−)	0.641_{9}	0.470_{9}	0.502_{9}
		5	0.692_{11}	0.587_{11}	0.640_{11}
	Random	1 (+)	0.691_{11}	0.570_{11}	0.658_{11}
		1 (−)	0.655_{10}	0.490_{10}	0.514_{12}
		5	0.694_{11}	0.582_{11}	0.653_{11}
word2vec		1 (+)	0.624	0.419	0.479
		1 (−)	0.661	0.506	0.559
		5	0.688	0.559	0.601

Table 4: Results of DIFFVEC on DEMELO and on CROWD using a single positive (1 (+)) or negative (1 (−)) $a_{ext} - a_{mild}$ pair, and five pairs (5).

coverage goes down to 11% on CROWD. This highlights the limitations of the approach, as well as the efficiency of our model which combines high performance and coverage.

5.3 Further Exploration of DIFFVEC

Given the high performance of the DIFFVEC method in the ranking task, we carry out additional experiments to explore the impact that the choice of scales and sentences has on the intensity vector quality. We test the method with a $\overrightarrow{dVec}$ vector built from a single $a_{ext} - a_{mild}$ pair of either positive (*awesome-good*) or negative (*horrible-bad*) polarity, that we respectively call DIFFVEC-1 (+)/(−). We also experiment with increasing the number of scales, adding *ancient-old*, *gorgeous-pretty* and *hideous-ugly* to form DIFFVEC-5. The scales are from WILKINSON, so we exclude this dataset from the evaluation.

Results are given in Table 4. We observe that a small number of word pairs is enough to build a $\overrightarrow{dVec}$ with competitive performance. Interestingly, DIFFVEC-1 (+) with random sentences obtains the best pairwise accuracy on DEMELO. The fact that the method performs so well with just a few pairs (instead of a whole dataset as in Table 3) is very encouraging, making our approach easily applicable to other datasets and languages.

A larger number of scales is beneficial for the method with static word2vec embeddings, which seem to better capture intensity on the negative scale. For BERT, instead, intensity modeled using a positive pair gives best results across the board. The use of five pairs of mixed polarity improves results over a single negative pair, and has comparable performance to the single positive one.

Finally, we compare the performance of DIFFVEC-1 (+)/(−) and DIFFVEC-5 when the contextualised representations are extracted from a single sentence instead of ten. Our main observation is that reducing the number of sentences harms performance, especially when the sentence used is randomly selected. Detailed results are included in Appendix D.

6 Indirect Question Answering

We conduct an additional evaluation in order to assess how useful DIFFVEC adjective rankings can be in a real application. As in Cocos et al. (2018), we address Indirect Question Answering (QA) (de Marneffe et al., 2010). The task consists in interpreting indirect answers to YES/NO questions involving scalar adjectives. These do not straightforwardly convey a YES or NO answer, but the intended reply can be inferred. For example, if someone is asked "*Was it a good ad?*" and replies "*It was a great ad*", the answer is YES. This makes Indirect QA a good fit for scalar adjective ranking evaluation since it allows to directly assess a model's capability to detect the difference in intensity and direction (positive or negative) in an adjective pair.

We use the de Marneffe et al. (2010) dataset for evaluation, which consists of 125 QA pairs manually annotated with their implied answers (YES or NO). We adopt a decision procedure similar to the one proposed by de Marneffe et al. (2010). We compute the BERT embeddings of the adjective in the question (a_q) and the adjective in the answer (a_a). If a_a (e.g., *great*) has the same or higher intensity than a_q (e.g., *good*) the prediction

		Method	Acc	P	R	F
BERT	ukWaC	DIFFVEC-1 $(+)_{10}$	0.715	0.677	0.692	0.685
		DIFFVEC-DM$_{12}$	0.707	0.670	0.689	0.678
		DIFFVEC-CD$_{12}$	0.675	0.635	0.648	0.642
		DIFFVEC-WK$_{11}$	**0.740**	**0.712**	**0.739**	**0.725**
	Flickr	DIFFVEC-1 $(+)_{9}$	0.699	0.663	0.680	0.672
		DIFFVEC-DM$_{11}$	0.699	0.659	0.673	0.666
		DIFFVEC-CD$_{10}$	0.691	0.653	0.667	0.660
		DIFFVEC-WK$_{5}$	0.683	0.646	0.661	0.654
	Random	DIFFVEC-1 $(+)_{9}$	0.715	0.677	0.692	0.685
		DIFFVEC-DM$_{10}$	0.724	0.691	0.713	0.702
		DIFFVEC-CD$_{12}$	0.667	0.629	0.642	0.636
		DIFFVEC-WK$_{11}$	0.699	0.667	0.688	0.677
word2vec		DIFFVEC-1 $(+)$	0.667	0.633	0.650	0.641
		DIFFVEC-DM	0.602	0.554	0.559	0.557
		DIFFVEC-CD	0.593	0.548	0.553	0.551
		DIFFVEC-WK	0.585	0.543	0.547	0.545
Baselines		FREQ	0.593	0.548	0.553	0.551
		SENSE	0.593	0.560	0.568	0.564
		MAJ	0.691	0.346	0.500	0.409
		$Previous_1$	0.610	0.597	0.594	0.596
		$Previous_2$	0.728	0.698	0.714	0.706
		$Previous_3$	0.642	0.710	0.683	0.684

Table 5: Results of our DIFFVEC method with contextualised (BERT) and static (word2vec) embeddings on the indirect QA task. We compare to the frequency, polysemy and majority baselines, and to results from previous work. $Previous_1$ stands for de Marneffe et al. (2010), $Previous_2$ for Kim and de Marneffe (2013) (the only result on 125 pairs), $Previous_3$ for Cocos et al. (2018).

is YES; otherwise, the prediction is NO. If the answer contains a negation, we switch YES to NO, and NO to YES. In previous work, indirect QA evaluation was performed on 123 or 125 examples, depending on whether cases labelled as "uncertain" were included (de Marneffe et al., 2010; Kim and de Marneffe, 2013; Cocos et al., 2018). We report all available results from previous work, and our scores on the 123 YES/NO examples as in the most recent work by Cocos et al. (2018). We report results using DIFFVEC with the adjustment for ties, where two adjectives are considered to be of the same intensity if they are similarly close to $\overrightarrow{dVec}$ ($diffsim$ = sim($\overrightarrow{dVec}$, $\overrightarrow{a_q}$) − sim($\overrightarrow{dVec}$, $\overrightarrow{a_a}$)). If the absolute value of $diffsim < 0.01$, we count them as a tie. We compare our method to previous results, to FREQ and SENSE, and to a baseline predicting always the majority label (YES). Results of this evaluation are given in Table 5. DIFFVEC with BERT embeddings outperforms the baselines and all previous approaches, and presents a clear advantage over DIFFVEC with static word2vec representations. Best performance is obtained when $\overrightarrow{dVec}$ is obtained from the Wilkinson dataset (DIFFVEC-WK). The $\overrightarrow{dVec}$ obtained from CROWD seems to be of lower quality. DIFFVEC-CD and DIFFVEC-DM improve over the baselines but do not achieve higher performance than the model of Kim and de Marneffe (2013).

7 Discussion

Our initial exploration of the knowledge encoded in BERT representations about scalar adjectives using BERTSIM (Section 5.1) showed they can successfully rank them by intensity. Then our DIFFVEC method (Sections 5.2 and 5.3) outperformed BERTSIM, providing even better ranking predictions with as few resources as a single adjective pair. This difference can be due to the composition of the vectors in the two cases. The a_{ext} representation in BERTSIM contains information about the meaning of the extreme adjective alongside its intensity, while the $\overrightarrow{dVec}$ vector is a cleaner representation of intensity: The subtraction of $\overrightarrow{a_{mild}}$ from $\overrightarrow{a_{ext}}$ removes the common core meaning expressed by their scale (e.g., BEAUTY, TEMPERATURE, SIZE). Consequently, $\overrightarrow{dVec}$ is a pure and general representation of intensity which can successfully serve to rank adjectives from any scale, as shown by our results. The DIFFVEC method can estimate adjectives' relative intensity on the fly, and performs better than the BERTSIM model which needs a reference point to propose a ranking. It does not use any external knowledge source – a requirement in previous approaches – and one of its highest performing variations (DIFFVEC-1 $(+)$) makes best quality predictions with a single adjective pair example.

Our assumption concerning the need for the sentences used for extracting BERT representations to be a good semantic fit for adjectives in a scale, has not been confirmed by our evaluation. Precisely, differences between our methods when relying on carefully vs randomly selected sentences are minor. This might be due to several reasons: One is that although BERT representations are contextualised, they also encode knowledge about the meaning and intensity of words acquired through pre-training, independent of the new context of use. Another possible explanation is that due to the skewed distribution of word senses (Kilgarriff, 2004; McCarthy et al., 2004), a high proportion of our randomly selected sentences might contain instances of the adjectives in their most frequent sense. If this is

also the meaning of the corresponding scale, then the sentences are a good fit.

The DIFFVEC-1 (+) method, which uses a vector derived from a single positive pair, yields consistently better results than DIFFVEC-1 (−) which relies on a single negative pair. To better understand this difference in performance, we examine the composition of DEMELO and CROWD, specifically whether there is an imbalance in terms of polarity as reflected in the frequency of positive vs negative adjectives in the two datasets. We check the polarity of the adjectives in two sentiment lexicons: SO-CAL (Taboada et al., 2011) and AFINN-165 (Nielsen, 2011). The two lexicons cover a portion of the adjectives in DEMELO and CROWD: 68% and 79%, respectively. The DEMELO dataset is well-balanced in terms of positive and negative adjectives: 51% and 49% of the covered adjectives fall in each category. In CROWD, we observe a slight skew towards positive: 61% vs 39%. According to this analysis, the difference in performance between the two methods could only partially be explained by an imbalance in terms of polarity.

We perform an additional analysis based on the Google Ngram frequency of the positive and negative words that were used for deriving DIFFVEC. The adjectives *good* (276M) and *awesome* (10M) are more frequent than *bad* (65M) and *horrible* (4M). In fact, we find that the 1,000 most frequent positive words in SO-CAL and AFINN are, on average, much more frequent (18M) than the 1,000 most frequent negative words (8M). Word frequency has a direct impact on word representations, since having access to sparse information about a word's usages does not allow the model to acquire rich information about its linguistic properties as in the case of frequent words. The high frequency of *good* and *awesome* results in better quality representations than the ones obtained for their antonyms, and could explain to some extent the improved performance of DIFFVEC-1 (+) compared to DIFFVEC-1 (−) with BERT embeddings. However, this analysis does not explain the difference in the performance of DIFFVEC (+) and (−) between BERT and word2vec. This would require a better understanding of how words with different polarity (antonyms) are represented in BERT's space compared to word2vec, and how negation affects their representations. We leave these explorations for future work.

Regarding the performance of different BERT layers, we observe that knowledge relevant for scalar adjective ranking is situated in the last layers of the Transformer network. Figure 3 shows how the performance of DIFFVEC-1 (+) changes across different BERT layers: model predictions improve after layer 3, and performance peaks in one of the last four layers. This is in accordance with the findings of Tenney et al. (2019) that semantic information is mainly located in the upper layers of the model, but is more spread across the network than syntactic information which is contained in a few middle layers.

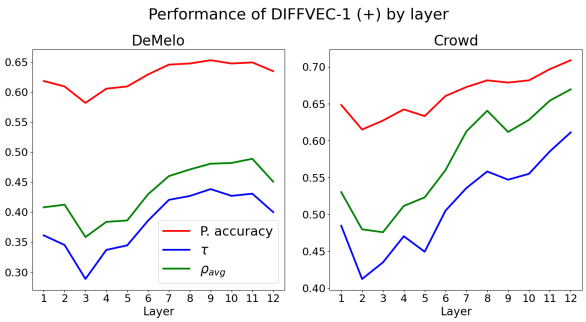

Figure 3: Performance of DIFFVEC-1 (+) with ukWaC sentences across BERT layers.

8 Conclusion

We have shown that BERT representations encode rich information about the intensity of scalar adjectives which can be efficiently used for their ranking. Although our method is simple and resource-light, solely relying on an intensity vector which can be derived from as few as a single example, it clearly outperforms previous work on the scalar adjective ranking and Indirect Question Answering tasks. Our performance analysis across BERT layers highlights that the lexical semantic knowledge needed for these tasks is mostly located in the higher layers of the BERT model.

In future work, we plan to extend our methodology to new languages, and experiment with multilingual and language specific BERT models. To create scalar adjective resources in new languages, we could either translate the English datasets or mine adjective scales from starred product reviews as in de Marneffe et al. (2010). Our intention is also to address adjective ranking in full scales (instead of half-scales) and evaluate the capability of contextualised representations to detect polarity.

Acknowledgements

This work has been supported by the French National Research Agency under project ANR-16-CE33-0013. The work is also part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 771113). We thank the reviewers for their thoughtful comments and valuable suggestions.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Barcelona, Spain.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. In *LDC2006T13*, Philadelphia, Pennsylvania. Linguistic Data Consortium.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Anne Cocos, Skyler Wharton, Ellie Pavlick, Marianna Apidianaki, and Chris Callison-Burch. 2018. Learning Scalar Adjective Intensity from Paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1752–1762, Brussels, Belgium. Association for Computational Linguistics.

Sunipa Dev and Jeff M Phillips. 2019. Attenuating Bias in Word Vectors. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Naha, Okinawa, Japan.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Bart Geurts. 2010. *Quantity implicatures*. Cambridge University Press.

Derek Gross and Katherine J Miller. 1990. Adjectives in WordNet. *International Journal of lexicography*, 3(4):265–277.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 172–182, Columbus, Ohio, USA. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher Kennedy and Louise McNally. 2005. Scale Structure and the Semantic Typology of Gradable Predicates. *Language*, 81:345–381.

Adam Kilgarriff. 2004. How Dominant Is the Commonest Sense of a Word? Lecture Notes in Computer Science (vol. 3206), Text, Speech and Dialogue, Sojka Petr, Kopeček Ivan, Pala Karel (eds.), pages 103–112. Springer, Berlin, Heidelberg.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving Adjectival Scales from Continuous Space Word Representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630, Seattle, Washington, USA. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What Substitutes Tell Us - Analysis of an "All-Words" Lexical Substitution Corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020. A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York City, NY, USA.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. "Was It Good? It Was Provocative." Learning the Meaning of Scalar Adjectives". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Uppsala, Sweden. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 279–286, Barcelona, Spain.

Louise McNally. 2016. Scalar alternatives and scalar inference involving adjectives: A comment on van Tiel, et al. 2016. In Ruth Kramer Jason Ostrove and Joseph Sabbagh, editors, *Asking the Right Questions: Essays in Honor of Sandra Chung*, pages 17–28.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Gerard de Melo and Mohit Bansal. 2013. Good, Great, Excellent: Global Inference of Semantic Intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint:1301.3781v3*.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC 2011 Workshop on 'Making Sense of Microposts: Big things come in small packages'*, volume 718 in CEUR Workshop Proceedings, pages 93–98.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Ajay Patel, Alexander Sands, Chris Callison-Burch, and Marianna Apidianaki. 2018. Magnitude: A Fast, Efficient Universal Vector Embedding Utility Package. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 120–126, Brussels, Belgium. Association for Computational Linguistics.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv preprint arXiv:2003.07082*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. *arXiv preprint arXiv:2004.07667*.

Sven Rill, J. vom Scheidt, Johannes Drescher, Oliver Schütz, Dirk Reinel, and Florian Wogenstein. 2012. A generic approach to generate opinion lists of phrases for opinion mining applications. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, pages 1–8, Beijing, China.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. *arXiv preprint:2002.12327v1*.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short*

Papers, pages 117–122, Gothenburg, Sweden. Association for Computational Linguistics.

Raksha Sharma, Mohit Gupta, Astha Agarwal, and Pushpak Bhattacharyya. 2015. Adjective Intensity and Sentiment Analysis. In *Proceedings of the 2015 Conference on Empirical Methods for Natural Language Processing*, pages 2520–2526, Lisbon, Portugal. Association for Computational Linguistics.

Vera Sheinman, Christiane Fellbaum, Isaac Julien, Peter Schulam, and Takenobu Tokunaga. 2013. Large, huge or gigantic? Identifying and encoding intensity relations among adjectives in WordNet. *Language resources and evaluation*, 47(3):797–816.

Vera Sheinman and Takenobu Tokunaga. 2009. AdjScales: Visualizing Differences between Adjectives for Language Learners. *IEICE Transactions on Information and Systems*, 92-D:1542–1550.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M. Lai. 2015. Corpus-based discovery of semantic intensity scales. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–493, Denver, Colorado. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. oLMpics – On what Language Model Pre-training Captures. arXiv preprint arXiv:1912.13283v1.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar Diversity. *Journal of semantics*, 33(1):137–175.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019b. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Bryan Wilkinson. 2017. *Identifying and Ordering Scalar Adjectives Using Lexical Substitution*. Ph.D. thesis, University of Maryland, Baltimore County.

Bryan Wilkinson and Tim Oates. 2016. A Gold Standard for Scalar Adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2669–2675, Portorož, Slovenia. European Language Resources Association (ELRA).

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint:1609.08144*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *Journal of General Psychology*, 33(2):251–256.

A Hearst Patterns

Figure 4 illustrates the dependency structure of the following Hearst patterns:

- *[NP] and other [NP]*
- *[NP] or other [NP]*
- *[NP] such as [NP]*
- *Such [NP] as [NP]*
- *[NP], including [NP]*
- *[NP], especially [NP]*
- *[NP] like [NP]*

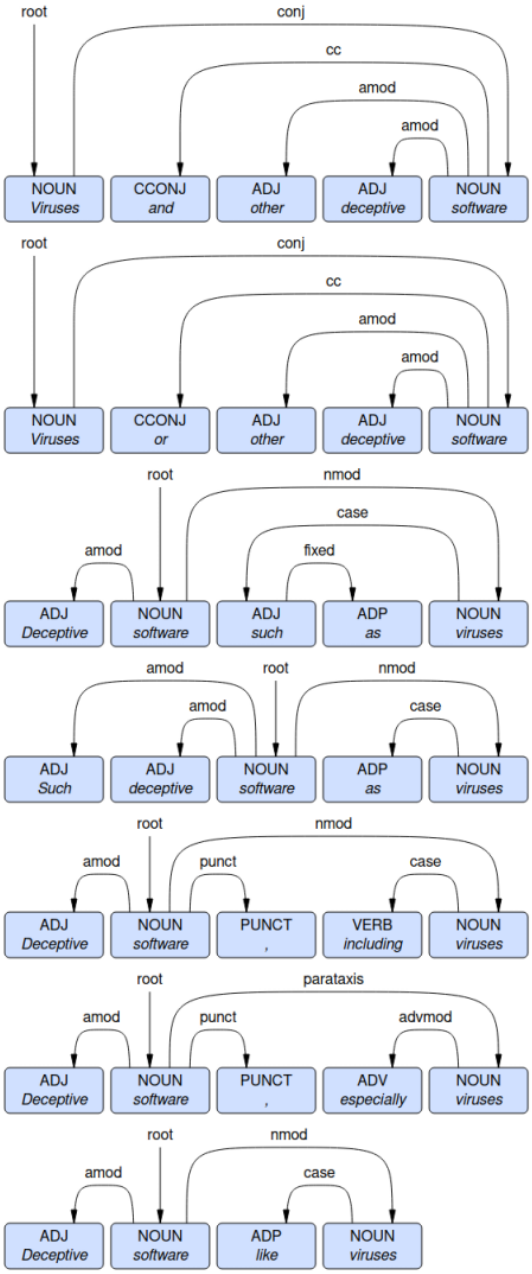

Figure 4: Dependency structure of Hearst patterns.

We use these patterns to detect sentences where adjective substitution should not take place, as described in Section 4.2 of the paper. We remove these sentences from our ukWaC and Flickr datasets.[17]

B Evaluation of Sentence Selection Methods

To identify the most appropriate method for selecting sentences where all adjectives in a scale fit, we use data from the Concepts in Context (CoInCo) corpus (Kremer et al., 2014). CoInCo contains sentences where content words have been manually annotated with substitutes which come with a frequency score indicating the number of annotators who proposed each substitute. We collect instances of adjectives, nouns and verbs in their base form.[18] For a word w, we form instance pairs (w_i-w_j with $i \neq j$) with similar meaning as reflected in their shared substitutes. We allow for up to two unique substitutes per instance, which we assign to the other instance in the pair with zero frequency. We keep instances with n substitutes, where $2 \leq n \leq 8$ (the lowest and highest number of adjectives in a scale). This results in 5,954 pairs.

We measure the variation in an instance pair in terms of substitutes using the *coefficient of variation* (VAR). VAR is the ratio of the standard deviation to the mean and is, therefore, independent from the unit used. A higher VAR indicates that not all substitutes are good choices in a context. We keep the 500 pairs with the highest VAR difference, where one sentence is a better fit for all substitutes than the other. For example, *private*, *individual* and *person* were proposed as substitutes for *personal* in "*personal insurance lines*", but *private* was the preferred choice for "*personal reasons*". The tested methods must identify which sentence in a pair is a better fit for all substitutes.

For sentence selection, we experiment with the three fluency calculation methods presented in Section 4.2: BERTPROB (the BERT probability of each substitute to be used in the place of the [MASK] token); BERTPPX (the perplexity assigned by BERT to the sentence generated through substitution); and CONTEXT2VEC (the cosine similarity between the context2vec representations of a substitute and the context).

We also test VAR and standard deviation (STD) as metrics for measuring variation in the fluency scores assigned to a sentence pair by the three methods. We evaluate the sentence selection methods and variation metrics on the 500 pairs retained from CoInCo. We report their accuracy, calculated as the proportion of pairs where a method correctly guesses the instance in a pair with the lowest variation. We compare results to those of a baseline that always proposes the first instance in a pair. The results in Table 6 show that the task is difficult for

[17] Graphs in Figure 4 were created with the visualisation tool available at https://urd2.let.rug.nl/~kleiweg/conllu/

[18] This filtering serves to control for morphological variation which could result in unnatural substitutions since CoInCo substitutes are in lemma form.

Method	Variation Metric	Accuracy
BERTPROB	STD	0.524
	VAR	0.488
BERTPPX	STD	0.518
	VAR	0.536
CONTEXT2VEC	STD	**0.594**
	VAR	0.588
1st sentence Baseline		0.506

Table 6: Accuracy of the three fluency calculation methods on the 500 sentence pairs collected from CoInCo. Comparison to a first sentence baseline.

all methods. Their accuracy is slightly higher than the baseline accuracy, which outperforms BERTPROB with VAR. The combination that gives best accuracy is CONTEXT2VEC with STD (0.594). We use this combination of metrics in our experiments.

Table 7 shows examples of sentences retained after this filtering for two adjective scales. CONTEXT2VEC tends to favour sentences where all adjectives in a scale fit well. We also give an example of a sentence randomly selected from ukWaC (Random) for a scale. These sentences usually reflect a frequent sense of a word in the scale.

C Adjustment for Ties

Table 8 contains results of the DIFFVEC method with the adjustment for ties. For two adjacent adjectives (a_i, a_j) in the ranking proposed by DIFFVEC, we check if their cosine similarities to $\overrightarrow{dVec}$ are very close ($diffsim$ = sim($\overrightarrow{dVec}$, $\overrightarrow{a_i}$) - sim($\overrightarrow{dVec}$, $\overrightarrow{a_j}$). If the absolute value of $diffsim < 0.01$, we count them as a tie, meaning that a_i and a_j are considered to be situated at the same intensity level. Note that this procedure may give different results when the pairwise comparison starts at different ends of the proposed ranking. We establish ties starting from the a with lowest intensity in the ranking proposed by DIFFVEC.

D DIFFVEC with a Single Sentence

Table 9 contains results for DIFFVEC-1 $(+)/(-)$ and DIFFVEC-5 when using a single sentence for building $\overrightarrow{dVec}$.

Scale: *wrong* → *immoral* → *sinful* → *evil*		
Method	Corpus	Sentences
context2vec-STD	ukWaC	I believe that war is *immoral*.
	Flickr	This boy was on the *wrong* end of this snowball fight.
Random	ukWaC	The author saw him and let him thru but not his mate as he had queued the *wrong* way.
Scale: *old* → *obsolete* || *outdated*		
Method	Corpus	Sentences
context2vec-STD	ukWaC	(...) Chekhov was misunderstood and frequently seen by critics as merely an irreverent recorder of an *obsolete* way of life (...)
	Flickr	Two preschool aged boys are looking at an *old* locomotive.
Random	ukWaC	(...) rustic dialogue and good *old* fashioned laughter (...)

Table 7: Examples of sentences from our SENT-SETs selected with the context2vec-STD method compared to sentences randomly selected from ukWaC.

			DEMELO (DM)			CROWD (CD)			WILKINSON (WK)		
		Method	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}
BERT	ukWaC	DIFFVEC-DM	-	-	-	0.733_{8}	0.673_{8}	$\mathbf{0.749}_{12}$	0.885_{6}	0.830_{11}	0.826_{6}
		DIFFVEC-CD	$\mathbf{0.644}_{8}$	$\mathbf{0.452}_{8}$	$\mathbf{0.518}_{8}$	-	-	-	0.820_{10}	0.721_{11}	0.780_{11}
		DIFFVEC-WK	0.546_{6}	0.295_{6}	0.324_{6}	0.721_{7}	0.627_{10}	0.698_{10}	-	-	-
	Flickr	DIFFVEC-DM	-	-	-	0.746_{12}	$\mathbf{0.685}_{12}$	0.718_{8}	$\mathbf{0.902}_{9}$	$\mathbf{0.851}_{9}$	$\mathbf{0.871}_{4}$
		DIFFVEC-CD	0.605_{11}	0.388_{11}	0.465_{11}	-	-	-	0.836_{8}	0.746_{7}	0.762_{7}
		DIFFVEC-WK	0.541_{2}	0.296_{1}	0.299_{1}	0.702_{8}	0.647_{8}	0.710_{8}	-	-	-
	Random	DIFFVEC-DM	-	-	-	$\mathbf{0.724}_{9}$	0.652_{9}	0.719_{8}	0.885_{11}	0.818_{6}	0.833_{10}
		DIFFVEC-CD	0.619_{8}	0.412_{8}	0.488_{8}	-	-	-	0.819_{12}	0.765_{10}	0.833_{10}
		DIFFVEC-WK	0.522_{2}	0.251_{6}	0.285_{6}	0.712_{10}	0.614_{9}	0.680_{9}	-	-	-
word2vec		DIFFVEC-DM	-	-	-	0.648	0.508	0.550	0.754	0.583	0.655
		DIFFVEC-CD	0.604	0.403	0.446	-	-	-	0.803	0.656	0.661
		DIFFVEC-WK	0.568	0.329	0.402	0.606	0.414	0.445	-	-	-

Table 8: Results of our DIFFVEC adjective ranking method on the DEMELO, CROWD and WILKINSON datasets with the adjustment for ties. We report results with contextualised (BERT) representations obtained from different SENT-SETs (ukWaC, Flickr, Random) and with static (word2vec) vectors.

			DEMELO			CROWD		
		# Scales	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}
BERT	ukWaC	1 (+)	0.651_{10}	0.433_{10}	0.501_{10}	0.682_{10}	0.553_{10}	0.622_{7}
		1 (−)	0.597_{1}	0.315_{1}	0.352_{1}	0.639_{12}	0.458_{12}	0.543_{12}
		5	$\mathbf{0.655}_{7}$	$\mathbf{0.443}_{7}$	$\mathbf{0.530}_{7}$	$\mathbf{0.691}_{11}$	$\mathbf{0.575}_{11}$	$\mathbf{0.675}_{11}$
	Flickr	1 (+)	0.639_{9}	0.410_{9}	0.432_{9}	0.676_{8}	0.550_{8}	0.604_{8}
		1 (−)	0.602_{3}	0.329_{3}	0.372_{3}	0.629_{4}	0.443_{4}	0.479_{4}
		5	0.624_{11}	0.380_{11}	0.452_{11}	0.683_{11}	0.562_{11}	0.606_{12}
	Random	1 (+)	0.631_{11}	0.401_{11}	0.451_{11}	0.676_{8}	0.536_{8}	0.589_{8}
		1 (−)	0.611_{9}	0.356_{9}	0.444_{9}	0.648_{11}	0.479_{11}	0.500_{11}
		5	0.622_{4}	0.371_{4}	0.417_{3}	0.685_{7}	0.559_{7}	0.588_{7}
word2vec		1 (+)	0.602	0.334	0.364	0.624	0.419	0.479
		1 (−)	0.613	0.359	0.412	0.661	0.506	0.559
		5	0.641	0.415	0.438	0.688	0.559	0.601

Table 9: Results of DIFFVEC using a single positive (1 (+)) or negative (1 (−)) adjective pair, and five pairs (5). These are results obtained with a $\overrightarrow{dVec}$ built from only one sentence (instead of ten in Table 4 of the paper).